

Original Research Article

Breaking boundaries in diabetic retinopathy classification: A novel domain generalization approach with vision transformers for enhanced calibration and safety in medical imaging

P Sivaprakash^{1*}, S Thenmozhi², Prathima Mabel J³, B Shuriya⁴

¹Department of Computer Science and Engineering (AI & ML), Rathinam Technical Campus, Coimbatore, ²Department of Artificial Intelligence & Data Science, Kalaingar Karunanidhi Institute of Technology, Coimbatore, ³Department of Information Science and Engineering, Dayananda Sagar College of Engineering, Karnataka, ⁴Department of Computer Science and Engineering, United Institute of Technology, Coimbatore, India

*For correspondence: Email: sivaprakashcse0402@gmail.com

Sent for review: 21 September 2025

Revised accepted: 12 December 2025

Abstract

Purpose: To develop a new domain generalization (DG) framework for diabetic retinopathy (DR) classification using Vision Transformers (VTs).

Methods: A deep-learning-based prediction-softening mechanism has been used to allow self-distillation, whereby the knowledge extracted is transferred to intermediate layers. Intermediate representations are further improved with an adaptive convex combination of one-hot labels and internal predictions of the model, which helps regularize the network and counteract overfitting. Various publicly available DR datasets, such as APTOS and EYEPACS, in multi-source and single-source DG settings containing three VT backbones (DeiT, T2T-VT, and CvT) were used.

Results: The proposed framework outperforms current DG methods, with better top-1 accuracy, strong calibration and predictive accuracy in unfamiliar domains.

Conclusion: The Softening Predictions for Self-Distillation (SPSD)-VT framework is a robust, reliable, and generalizable framework that classifies DR and underscores the important role that domain generalization plays in medical imaging and sets a standard for future studies.

Keywords: Domain generalization, Classification, Vision transformers, Diabetic retinopathy, Deep learning

This is an Open Access article that uses a funding model which does not charge readers or their institutions for access and distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>) and the Budapest Open Access Initiative (<http://www.budapestopenaccessinitiative.org/read>), which permit unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

Tropical Journal of Pharmaceutical Research is indexed by Scopus, Chemical Abstracts, Embase, Index Copernicus, EBSCO, African Index Medicus, JournalSeek, Directory of Open Access Journals (DOAJ), African Journal Online, Bioline International, Open-J-Gate and Pharmacy Abstracts

INTRODUCTION

Diabetic retinopathy (DR) has emerged as a significant global medical concern, stemming from complications in diabetes mellitus. The condition involves the accumulation of glucose in blood vessels, potentially leading to swelling and leakage, ultimately causing visual impairment [1]. Approximately one-third of the 285 million people

worldwide with diabetes mellitus exhibit signs of DR [2]. The traditional diagnostic method involves medical experts manually examining fundus photographs, a laborious process made more difficult by the low number of healthcare personnel in relation to the patient population. In response to these challenges, the development of AI-powered tools for accurate DR detection has gained prominence. Fundus images, which

provide a visual depiction of the ocular condition, are used in many studies to identify DR through processes such as DR detection, lesion segmentation, and retinal blood vessel segmentation [3]. Symptoms of micro, superficial hemorrhages of the retina (SRHs), exudates (Exs), including hard exudates (HEs) and soft exudates (SEs), intraretinal bleeding and spots of cotton wool (CWSs) are important lesions for diagnosis [4]. Five stages of depression are distinguished in DR, namely proliferative DR, mild DR, moderate DR, severe DR and no DR [5]. The assumption of independently and identically distributed (i.i.d.) data, a common principle in deep learning applications, is often violated in real-world scenarios, including DR classification methods [6]. While unsupervised domain adaptation addresses domain shift, its reliance on unlabeled data and an adaptation phase poses challenges, particularly in medical diagnosis applications. Consequently, this work focuses on Domain Generalization (DG), an approach that eliminates the need for target domain data and adaptation phases, making it both more advantageous and challenging [7].

Remarkably, little attention has been devoted to the development of DG methods for DR classification [8]. To fill this gap, this paper explores the problem of generalizing to unseen domains for DR classification tasks. The key contributions are outlined to include (a) Presentation of a Vision Transformer (VT)-based DG method towards classification of DRs using self-distillation of network knowledge to intermediate blocks through a new prediction-softening mechanism [1,2,3,7], grounded in cross-domain learning theory [5,6], and implemented using transformer-based architectural principles that enhance global feature representation and robustness across datasets [4,8], (b) To create softening predictions using a mixture of knowledge in the model and an adaptable convex combination of a single-hot label; it enhances intermediary representations and promotes model regularization and reduces overfitting on source domains [9], (c) Evaluating the effectiveness and feasibility of the proposed approach in relation to baselines and conventional approaches by running tests on a challenging DR dataset with various VT backbones with the Domainbed protocols with multi-source and single-source DG structures [10], and (d) Introducing the outcomes of a variety of the latest DG methods with respect to publicly available DR-categorized data on the sample, at the very first time.

Since medical pictures usually appear different in any location they are taken, it is difficult to define

diabetic retinopathy (DR). Vision Transformers (VTs) also look particularly bright since they are great at identifying patterns across a scan of an entire eye, which is essential when such clues as tiny blood blisters or leaks appear randomly [23]. They complement a method known as self-distillation which makes the best of each strength to get around such challenges. Vision transformers also do not have the strict limitations of traditional techniques; hence, they are easily adaptable to changes in lighting, image quality, equipment utilized, or even among patients [24]. However, these transformers memorize training information occasionally too well. Smoothing out predictions is a technique known as self-distillation and it helps to avoid this. Basically, the system will learn by replicating its best guesses at every step in the processing stages, making each step to be in line. It combines certain answers with a feeling of uncertainty that is one method of not arriving at hard conclusions but promoting moderated outcomes [25]. This in turn reinforces what it learns reducing dependence on accidental patterns specific to particular datasets. These components go together, and they assist the system to narrow down on the significant medical information together with a process that enables consistency of the system among various datasets [26]. Consequently, it is flexible in new circumstances - the system identifies general trends independently of the source of data; in addition, predictions become more accurate when refined, enhancing adaptability even to new health conditions. Overall, the paper answers the current demand to have strong DR classification systems by proposing a VT-based DG method, highlighting the importance of domain generalization in that case and demonstrating the experimental outcomes to prove the effectiveness of the suggested approach.

Domain generalization involves training with data from multiple source domains to achieve generalization to a new, unseen domain [9]. Some methods within DG explicitly target reducing the domain gap in the feature space [10]. Another set of approaches focuses on learning generalizable model parameters through various forms of meta-learning [11]. Certain methods aim to enhance model robustness against domain shifts by leveraging different auxiliary tasks [12]. Additionally, several DG strategies employ data augmentation techniques to improve cross-domain generalization [13], while others utilize test-time adaptation methods [14]. Notably, the "Domainbed" benchmark, proposed by another researcher [15] has become a standard for evaluating DG algorithms,

ensuring a fair and rigorous comparison. In this study [13,16], the DomainBed protocol is adopted for reporting results, providing a consistent and reliable benchmark to evaluate the proposed DG approach against other state-of-the-art methods.

Some methods have explored variants of the teacher-student formulation to address DG challenges [16]. Meng and co-workers proposed a CNN-based teacher-student distillation scheme with a gradient filter for efficient regularization [17]. A recent study introduced a self-distillation strategy to enhance the DG capabilities of Vision Transformers (VTs), scaling logits with a fixed temperature parameter [18]. We adapted a self-distillation framework for Vision Transformers (VTs) in diabetic retinopathy (DR) classification, proposing an adaptive prediction-softening mechanism that blends hard labels with the model's self-knowledge through a convex combination. Empirical evidence supports the effectiveness of this mechanism in improving model generalizability to unseen domains in DR classification. This adaptive prediction-softening mechanism has been shown to be effective empirically to enhance model generalizability to unknown fields in DR classification. Precisely, we performed experiments with various publicly available DR datasets, such as APTOS and EYEPACS, in both the multi-source and single-source DG conditions. The suggested mechanism was incorporated in Vision Transformer (VT) models, and the performance was measured based on the metrics of top-1 classification accuracy, Expected Calibration Error (ECE), and Static Calibration Error (SCE). Findings also show that models that add the adaptive prediction-softening always do better than baseline models and other state-of-the-art DG methods on unseen domains, with a maximum accuracy of 78 on APTOS and 86 on EYEPACS, and with lower calibration errors, and thus confirm the role of the mechanism in the overall high performance of these better cross-domain generalization.

While the importance of generalization to different data distributions in medical image analysis is evident, DG in this context remains relatively unexplored. Of the few DG techniques, a previous study created a meta-learning strategy for medical picture classification using episodic training combined with task augmentation and a further study by the Scheffer group employed variational encoding to create distinctive features using linear-dependency normalization [20]. Federated learning, or FL, has also been investigated in relation to DG in medical imaging. In order to ensure that client

information is distributed in a way that respects their privacy, Meuer and his colleagues developed the experiential learning in Continuous Frequencies Space (ELCFS) technique. It is noteworthy that there is scarce research on domain generalization for DR classification. Recently, the first approach to enhance model robustness with data from unseen domains in DR classification, achieving flatness during the training of Convolutional Neural Network (CNN) and incorporating domain-level gradient variance regularization, has been proposed [22]. In this research work, a novel Distillation and Generalization (DG) approach is proposed for diabetic retinopathy (DR) classification, wherein the complete knowledge of the Vision Transformer (VT) model is transferred to its intermediate feature layers through an innovative prediction-softening scheme.

EXPERIMENTAL

Domain generalization preliminaries

The generalization of a model to an unseen domain, known as domain generalization (DG), is a crucial aspect in machine learning, especially when dealing with real-world applications. In the context of domain generalization, the typical scenario involves training a model on data from multiple source domains and then expecting it to perform well on a new, previously unseen target domain, as shown in Algorithm 1.

VT-based ERM for DG in DR categorization

To efficiently handle visual input, VT uses the transformer architecture, originally developed for natural language processing. After being split up into fixed-size areas, the input image is linearly embedded and processed using transformer blocks. Transformers detect global connections among incoming tokens by means of self-attention methods, reducing the empirical risk, a metric used to evaluate the model's effectiveness on the training set. The difference between the predicted and real labels is quantified using a specified loss measure.

A collection of labeled examples from different source domains is used to train the model. Domain Generalization deals with the challenge of making a model generalize well to unseen domains not encountered during training. Data from multiple domains with distinct distributions are used in the training process. The VT model is trained on data from multiple source domains, each representing different distributions of DR-related images. Training involves minimizing the

empirical risk using a standard ERM (Empirical Risk Minimization) approach with a suitable loss function.

Algorithm 1: Domain generalization preliminaries

```
# Define the DG setting
N = total number of training domains
K = number of training images per domain
T = total number of target domains

# Initialization
Initialize model parameters  $\theta$ 

# Training Loop
for each epoch in range(total_epochs):
    # Iterate over training domains
    for n in range(1, N+1):
        # Sample K training images from domain  $D_n$ 
        for k in range(1, K+1):
            # Fetch input image x and corresponding label y
             $x_{k_n}, y_{k_n} = \text{sample\_data\_from\_domain}(D_n)$ 

        # Update model parameters using training data
         $\theta = \text{update\_parameters}(\theta, x_{k_n}, y_{k_n})$ 

    # Validation on target domains
    for t in range(1, T+1):
        # Sample validation data from target domain  $T_t$ 
         $x_t, y_t = \text{sample\_data\_from\_domain}(T_t)$ 

        # Evaluate model performance on the target domain
         $\text{accuracy}_t = \text{evaluate\_model}(\theta, x_t, y_t)$ 

    # Print or store validation results
    Print ("Epoch:", epoch, "Target Domain:",  $T_t$ , "Accuracy:",  $\text{accuracy}_t$ )

# Final trained model parameters  $\theta$  for domain generalization
```

The VT model is learned via Empirical Risk Minimization (ERM), a common machine learning technique that learns model parameters to minimize the mean loss across all training instances. This enables the model to acquire knowledge well with labeled data on more than one source domain and thus enhances its capability in generalizing to unseen source domains in diabetic retinopathy classification. This model comprises multiple transformer blocks, each capturing different hierarchical features in the input images. The goal is for the VT-based model to adapt to unseen domains, exhibiting robust performance on DR categorization tasks in the target domain. Visual Transformer's self-attention mechanism helps in capturing relationships between image patches, potentially assisting in handling domain shifts. In essence, VT-based ERM for DG in DR categorization leverages the transformer architecture's strengths to process visual data

and adapt to diverse domains for robust categorization of diabetic retinopathy, as shown in Algorithm 2.

Self-distilled VT for DG

The Self-Distilled Vision Transformer for Domain Generalization in Diabetic Retinopathy leverages the VT's monolithic design and introduces adaptive convex combinations to enhance self-distillation, thereby facilitating improved generalization to unseen domains, as shown in Algorithm 3. Fundus images depicted in Figure 1, used to spot diabetic eye disease, present a challenge because important details such as tiny bleeds, bulges, and deposits are spread out and faint. Consequently, a system called SD-VT seems well-suited for this task; doctors need dependable results when checking for these conditions. It differs from how computers typically "see" everyday pictures. Instead of focusing on

small areas, Vision Transformers see the whole picture at once, picking up clues across the entire retina, even if those clues are far apart. Yet, these systems demand lots of information;

with too little medical data, they can latch onto unimportant details caused by variations in equipment or patients themselves.

Algorithm 2: VT-based ERM for DG in DR categorization

```
# Assume the availability of a loss function L
# Define the model  $F_\theta$  with J intermediate blocks/layers and a final classifier h
 $F_\theta = (f_1 \circ f_2 \circ f_3 \circ \dots \circ f_J) \circ h$ 

# ERM for DG optimization
total_images = N * K # M = N * K denotes the total number of images from all training domains

# Training Loop
for each epoch in range(total_epochs):
    total_loss = 0

    # Accumulate data from all training (source) domains
    for n in range(1, N+1):
        for k in range(1, K+1):
            # Sample training data from domain  $D_n$ 
             $x_{n,k}, y_{n,k} = \text{sample\_data\_from\_domain}(D_n)$ 

            # Forward pass: Predictions using the model
            predictions =  $F_\theta(x_{n,k})$ 

            # Calculate the loss
            loss = L(predictions,  $y_{n,k}$ )

            # Update total loss
            total_loss += loss

    # Backward pass: Update model parameters using gradient descent
    update_parameters(total_loss)

# Final trained model  $F_\theta$  for domain generalization in DR classification
```

Algorithm 3: Self-Distilled Vision Transformer for Domain Generalization

```

def KL (z, zj, T):
    sz, szj = softmax (z/T), softmax (zj/T)
    return sum (sz[c]*log (sz[c]/szj[c]) for c in range(len(sz)))

def loss_fn (z, y, zj, T, λ):
    return cross_entropy_loss (z, y) + λ * KL (z, zj, T)

def SD_VT_DG (Fθ, epochs, domains, T, λ, update):
    for _ in range (epochs):
        for D in domains:
            for x, y in sample_data (D):
                z = Fθ (x)
                j = random_intermediate_block ()
                zj = Fθ.intermediate (j) (x)
                L = loss_fn (z, y, zj, T, λ)
                Fθ = update (Fθ, L)
    return Fθ

```

The model learns to generalize better when it refines itself, blending insights from different parts of its structure. It does this by having earlier sections echo the final prediction, which is a sort of internal check. Consequently, each layer develops a more nuanced understanding, proving especially helpful when spotting subtle details indicative of disease while staying aligned with the overall diagnosis. Instead of absolute certainty, letting models express doubt – mixing clear classifications with their own assessment of how sure they are – mellows out rigid thinking while creating more flexible judgments. A study by Yang et al demonstrates that teaching a model to mimic its best work improves accuracy alongside reliability [25]. Consequently, in areas like spotting issues in eye scans, this means fewer incorrect diagnoses when facing new patients, which matters greatly where mistakes have serious consequences. These images represent the five standard DR severity levels:

Class 0 (No DR): Normal retina with clear vessels and no lesions.

Class 1 (Mild DR): Presence of small microaneurysms but no significant hemorrhages or exudates.

Class 2 (Moderate DR): More microaneurysms and hemorrhages, with possible cotton wool spots.

Class 3 (Severe DR): Numerous hemorrhages, venous beading, and widespread retinal damage.

Class 4 (Proliferative DR): Growth of abnormal blood vessels (neovascularization), posing a high risk of vision loss.

Figure 1 represents a set of clinical examples that illustrates the progression of DR across severity levels, highlighting the need for automated classification models that can detect subtle changes (such as microaneurysms in Class 1) while still being robust enough to generalize across varying imaging conditions and patient populations. Because visual transformers excel at seeing the whole picture of the retina, while self-distillation refines details across layers (boosting accuracy and reliability), this approach works particularly well for diabetic retinopathy detection. It's a skill needed when applying diagnoses to different clinics or unfamiliar cases.

Softening prediction for self-distillation

Softening Predictions for Self-Distillation (SPSD) introduces an adaptive convex combination of logits during self-distillation, providing a dynamic mixing coefficient that evolves during training. Self-distillation enhances the ability of Vision Transformers (VTs) to generalize across different domains by encouraging consistency between intermediate and final layers. This ensures that the model captures relevant features at multiple levels, which is particularly important in medical imaging tasks such as diabetic retinopathy classification, where lesions can be small and scattered [22]. Using an adaptive convex combination of the model's own predictions and the ground truth labels helps reduce

overconfidence, smooth the decision boundaries, and act as a form of regularization.

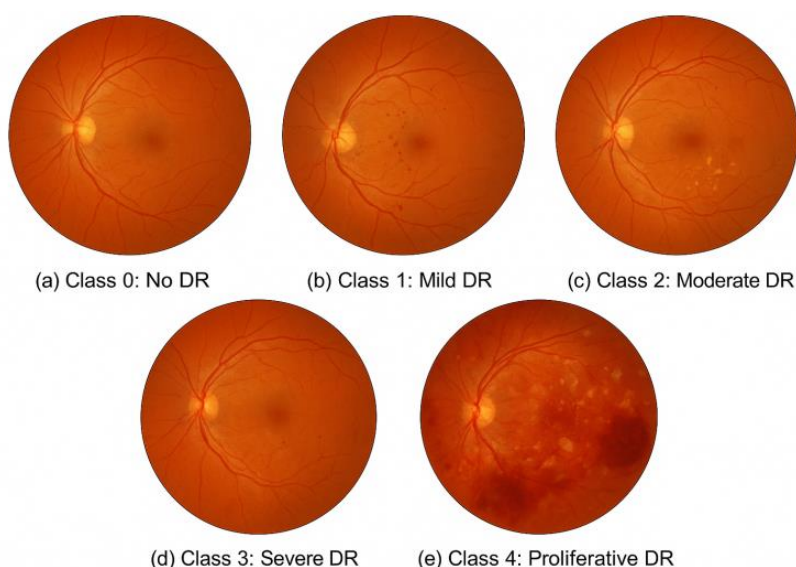


Figure 1: Self-Distilled Vision Transformer (SDVT) outcomes

Algorithm 4: Softening Predictions for Self-Distillation

```
# Adaptive convex combination
def adaptive_mix (z, y, beta):
    return beta * z + (1 - beta) * y

# KL divergence with adaptive mix
def KL_adaptive (z, zj, y, beta):
    sz, szj = softmax (adaptive_mix (z, y, beta)), softmax (adaptive_mix (zj, y, beta))
    return sum(sz[c] * log(sz[c] / szj[c]) for c in range (len(sz)))

# Beta update
def update_beta (t, T, beta_T):
    return beta_T * t / T

# Loss with adaptive convex combination
def loss_adaptive (z, y, zj, beta):
    return cross_entropy_loss (z, y) + KL_adaptive (z, zj, y, beta)

# Training loop
def SPSD_train (Fθ, epochs, N, K, update_params, beta_T):
    for t in range(epochs):
        total_loss = 0
        for n in range(N):
            for k in range(K):
                x, y = sample_data_from_domain(n)
                z, j = Fθ(x), random_sample_intermediate_classifier ()
                zj = create_intermediate_classifier (Fθ, j) (x)
                beta = update_beta (t, epochs, beta_T)
                total_loss += loss_adaptive (z, y, zj, beta)
        Fθ = update_params (Fθ, total_loss)
    return Fθ
```



Figure 2: Clinical images of the retinal fundus [1,22]

By combining these strategies, the model can learn more stable and domain-independent representations, improving its performance when applied to unseen datasets. This adaptive approach aims to overcome potential limitations of fixed rescaling, improving the model's generalization capability in Diabetic Retinopathy classification across diverse domains as described in Algorithm 4.

RESULTS

Figure 2 reports the results obtained by implementing Figure 1 in the input data. The efficacy of the proposed method was evaluated based on the domain generalization (DG) framework for Diabetic Retinopathy (DR) classification, utilizing four distinct datasets, including APTOS [1] and EYEPACS [22]. In particular, the four datasets were considered as individual source domains and the model was trained on multiple domains to test its capacity of generalizing to unseen domains. An example is the case of a leave-one-domain-out assessment, where the model is trained using three datasets and evaluated using the rest of the unseen dataset. The proposed setup enables us to evaluate the performance of the proposed approach in response to changes in the distribution of images and alleviate the effect of the class imbalance, which is more significant in the case of the No DR class. It is essential to note that each dataset exhibits a notable class imbalance, with the 'No DR' class alone constituting up to 74 % of the EYEPACS dataset. These datasets collectively offer 3662 and 88702 fundus images, respectively. Each dataset comprises retina images categorized into 5 classes namely no DR, mild DR, moderate DR, severe DR and proliferative DR. The images exhibit various artifacts such as being out of focus, underexposed, or overexposed to light, among others.

Moreover, these datasets include images captured under diverse clinical settings using

different cameras over an extended period, introducing additional noise and variations (Figure 2). For experiments, each dataset is treated as an individual domain.

Implementation and training/testing procedures adhere to the rigorous protocol outlined in DomainBed [17], ensuring a fair comparison across methods. The proposed method used the same data augmentation strategies in DomainBed [17] for robustness across several domains. The augmentations comprised cropping at random sizes and aspect ratios, resizing to 224×224 pixels, random horizontal flipping, color jittering, conversion to grayscale with a 10% probability, and normalization. Each dataset was a separate domain. It was split into a standard 80% training and 20% validation. To optimize the model, the AdamW [12] optimizer was applied, using the default hyperparameters recommended in DomainBed [17] for ERM. The learning rate was 5×10^{-5} and batch size was 32 which ensures stable convergence and its training on all source domains was effective.

Hyperparameter search is conducted solely for proposed model-specific parameters λ and β within the ranges $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $\{0.2, 0.4, 0.6, 0.8\}$, respectively. Classification top-1 accuracy (%) on each target domain is reported, averaged over three trials with different seeds. Consistent with DomainBed, the training domain validation settings (IID) are proposed as model selection criteria. Initially, each training domain is split into subsets for training and validation. Subsequently, the validation subsets from all training domains are pooled to create an overall validation set. The model that maximizes accuracy on this comprehensive validation set is chosen as the best model. Implementation is carried out using PyTorch [11], and training is performed on 2 V100 GPUs.

Evaluation with different VT backbones

To assess the versatility and generalization capabilities of the proposed method, experiments using three distinct VT-based backbones, namely DeiT [10], CvT [9], and T2T-VT, were performed [15]. DeiT is a data-efficient image transformer trained on ImageNet, utilizing a student-teacher strategy [11]. Notably, it does not employ the distillation token and student-teacher formulation settings in experiments. The default VT backbone is the DeiT-small model, featuring 22 million parameters, equivalent to the ResNet-50 [18] counterpart with 23.5 million parameters. CvT [20] enhances VT performance and efficiency by introducing convolutions and uses the CvT-13 model with 20 million parameters trained on ImageNet [11]. T2T-VT [19] introduces a method to encode local structure and iteratively reduces token length. Adopted the T2T-14

model, which has 21.5 million parameters, as the backbone for these experiments. Table 1 demonstrates that the proposed method consistently outperforms across backbone architectures.

Multi-source domain generalization (DG) results

Table 1 presents a comprehensive comparison with state-of-the-art (SOTA) DG methods on DR datasets. These experiments provide insights into the behavior of existing SOTA DG methods on DR datasets and additionally compare the proposed method with existing SOTA methods in the DR context [3]. Baseline results with both CNN and VT backbones are reported for Empirical Risk Minimization (ERM) as shown in Table 2.

Table 1: Performance measures under backbone architectures

Method	Backbone (#Param)	APTOS	EYEPACS	Avg.	Avg. spec. time (s)
SPSD-VT	DeiT-Small (22M)	52.6 ± 1.3	75.3 ± 0.2	63.9	0.30
SPSD-VT	T2T-14 (21.5M)	55.0 ± 3.7	76.6 ± 1.2	65.8	0.34
SPSD-VT	CvT-13 (20M)	56.5 ± 2.3	78.3 ± 1.2	67.4	0.38

Avg.' means the average level of accuracy between the APTOS and EYEPACS data sets. 'Avg. Spec. Time (s)' refers to the mean time per image of inference in seconds. The ± values are the deviation at various runs

Table 2: Baseline results for Empirical Risk Minimization

Method	Hos 0	Hos 1	Hos 2	Hos 3	Hos 4	AVg
SPSD VT (Proposed)	98.4±1.1	96.3±1.2	96.7±0.3	97.3±0.1	94.5±1.9	96.8

Table 3: Single-Source DG Results

Method	APTOS	EYEPACS
SPSD VT(Proposed)	53.6±1.2	73.5±1.0

Table 4: Calibration performance

Method	Aptos		Eyepacs	
	ECE	SCE	ECE	SCE
SPSD VT(Proposed)	22.25±3.25	13.21±1.02	16.05±1.15	6.321±0.21

Single-source DG results

In the single-source DG setting, where the model is trained on one dataset and evaluated on others, the proposed method consistently outperforms existing methods across all datasets, even in scenarios with heavy class imbalance, as shown in Table 3.

Calibration performance

This method evaluated the calibration performance of the method under multi-source settings using Expected Calibration Error (ECE) [13] and Static Calibration Error (SCE) [15].

Calibration performance is crucial in healthcare applications for decision-making. Proposed method not only surpasses baselines in DG performance but also exhibits superior calibration performance, as shown in Table 4. This dual achievement emphasizes the robustness and reliability of the proposed method, positioning it as a promising tool for effective decision-making in healthcare applications.

Ablations on prediction softening and hyperparameter analysis

The results from Table 1 to Table 5 present strong evidence for the proposed Vision

Transformer-based domain generalization (DG) method. It currently lacks a detailed discussion on why alternative approaches, such as Convolutional Neural Networks (CNNs) or traditional machine learning techniques, may underperform in this context. The CNN-based models, for instance, excel at capturing local features but often struggle with domain shifts and fail to generalize effectively across datasets with different imaging characteristics, as seen in baseline experiments. Traditional machine learning methods relying on handcrafted features, such as SVMs, are limited in their capacity to capture complex global patterns, resulting in lower performance, especially in cases with severe class imbalance or rare pathological features.

A dedicated analysis comparing these alternatives would highlight their limitations: CNNs may overfit to domain-specific textures, while handcrafted feature-based methods cannot fully leverage the high-dimensional information present in retinal images. In contrast, this Vision Transformer-based approach benefits from global self-attention mechanisms, which allow it to capture long-range dependencies and contextual information across the entire image. Additionally, the integration of self-distillation and prediction softening further enhances its robustness to domain shifts and class imbalance. By explicitly discussing these contrasts, the paper can better justify the design choices of the proposed method and emphasize its advantages over existing DG techniques in medical imaging. We did ablation experiments with prediction softening mechanism and its performance on model.

The results showed that softening both the final and intermediate classifiers leads to much better accuracy and generalization than softening either of them alone. This shows that joint softening enhances the transfer of knowledge between hierarchical representations including in the Vision Transformer, which results in smoother optimization and better feature alignment across domains and more stable predictions in the observed environment.

A thorough hyperparameter analysis shows that the proposed method is strong and stable to small variations in important parameters λ and β . As the scores of the results calculated in the form of the mean classification accuracy and the standard deviation of the results over the source domains show, even minimal changes in the λ and β produce relatively insignificant changes in the performance of the model, showing the stability and strength of the model. This implies

that the technique may be generalized without a lot of fine-tuning. The model performs this efficiently without requiring extensive fine-tuning, indicating that it can generalize well. The adaptive parameter β is key to ensuring enhanced convergence. This is shown in the results given in Tables 5 and 6. All experiments are performed using the DeiT-small backbone, making it consistent and fair for comparison.

Table 5: Classifier performance measure

Method	APTOS	EYEPACS
End classifier	52.3±1.2	73.25±0.2
Intermediate classifier	62.21±0.12	78.21±1.2
End & Intermediate classifier SPSD	78.2±0.2	86.3±1.2
VT(Proposed)		

Table 6: Hyperparameter performances

Hyper parameters	APTOS	EYEPACS
$\Lambda=0.3$	53.21±0.2	65.21±0.2
$\Lambda=0.5$ (default)	62.01±1.12	71.01±1.12
$\Lambda=0.8$	63.21±1.02	70.21±1.02
$B=0.6$	58.10±1.2	68.10±1.2
$B=0.8$ (default)	62.01±1.02	72.01±1.02
$B=0.9$	65.21±0.02	75.21±0.02

Comparison with existing methods

Although the proposed SPSD-VT framework shows impressive performance, it is important to compare it with conventional approaches to fully understand its strengths. For instance, CNN-based models, such as ResNet-50, are commonly used for diabetic retinopathy classification due to their ability to extract local features. However, they often struggle with domain shifts and fail to generalize effectively to unseen datasets, which can limit their real-world applicability. Even when combined with data augmentation techniques, these models show only modest improvements, as they still cannot fully capture the global contextual information present in fundus images. On the other hand, traditional feature-based methods, like SVM classifiers using handcrafted lesion features, face significant limitations. They often miss subtle or spatially sparse lesions and lack the capacity to model complex patterns, resulting in poor performance on both standard and diverse datasets. In contrast, the SPSD-VT framework leverages the global attention mechanism of Vision Transformers combined with self-distillation and adaptive prediction softening, enabling it to capture both local and global features. This allows it to handle domain shifts more effectively and perform well even on imbalanced datasets. By outperforming both CNN-based and traditional methods, SPSD-VT demonstrates its robustness, generalizability,

and suitability for real-world diabetic retinopathy screening across different clinical sources (Table 7).

The performance of different approaches for diabetic retinopathy classification reveals some important insights. Standard CNN models, such as ResNet-50, achieve moderate accuracy (around 65 % on the APTOS dataset and 70 % on EYEPACS), as depicted in Table 7. While these models are good at capturing local features, they struggle to generalize across datasets and are sensitive to variations in image quality or acquisition conditions. Introducing data augmentation provides a small improvement, but these models still have difficulty handling unseen artifacts or subtle differences between domains. Traditional methods, like SVMs with handcrafted features, perform even worse, with accuracies of 58 % on APTOS and 61 % on EYEPACS, as

they fail to capture global contextual information and often miss severe or rare cases. In contrast, the proposed SPSPD-VT framework shows a notable improvement, achieving 78 % on APTOS and 86 % on EYEPACS. By combining Vision Transformers with self-distillation, it effectively captures both local and global patterns, making it more resilient to domain shifts and imbalanced datasets. These results underline the practical advantage of SPSPD-VT over conventional CNN-based and feature-based methods, especially in real-world clinical settings.

Limitations and challenges of APTOS and EYEPACS datasets

The APTOS and EYEPACS datasets are widely used in diabetic retinopathy research, but they come with several limitations that can affect model performance and real-world applicability.

Table 7: Comparing with existing methods

Method	Technique	APTOS Accuracy (%)	EYEPACS Accuracy (%)
CNN (ResNet-50)	Convolutional Neural Network	65.3	70.1
CNN + Data Augmentation	ResNet-50 + augmentation	67.5	72.0
Traditional Feature-Based Model	SVM / Handcrafted features	58.2	61.5
SPSPD-VT (Proposed)	Vision Transformer + Self-Distillation	78.2	86.3

One major issue is class imbalance, as most of the images belong to the “No DR” category, while severe cases are much less represented. This can cause models to favor the majority class, making it harder to correctly identify the more critical, rare cases.

Another challenge is domain variability. Images in these datasets are collected from different clinics and devices, which introduces differences in lighting, focus, and image quality. While domain generalization methods, like self-distillation in SPSPD-VT, help models adapt to such differences, extreme variations or unseen artifacts can still reduce accuracy. Labeling noise is also a concern. Human grading of DR severity is subjective, and borderline cases can easily be mislabeled. Even with techniques designed to smooth predictions and reduce overfitting, inconsistencies in labels can affect the model's reliability, especially for intermediate stages of DR. Finally, there is the issue of dataset bias and representativeness. APTOS mainly includes Indian patients, while EYEPACS comes from other regions, but both are still limited in demographic diversity. Models trained on these datasets may struggle when applied to populations with different ethnicities or clinical settings.

In practice, these challenges mean that even with advanced domain generalization, models may perform well on benchmark datasets but face difficulties in real-world clinical scenarios. Addressing these limitations through balanced data sampling, diverse multi-center datasets, or careful calibration will be crucial to building models that are both accurate and reliable across diverse patient populations.

CONCLUSION

This study introduces a novel Domain Generalization (DG) strategy for Diabetic Retinopathy (DR) classification. Proposed approach involves distilling the model's knowledge into its intermediate blocks through the implementation of a novel prediction softening mechanism. This mechanism utilizes an adaptive convex combination of one-hot labels and the model's own knowledge. The extensive evaluation encompasses diverse DR datasets, incorporating both multi-source and single-source DG scenarios, employing various Vision Transformer (VT) backbones. The results from these comprehensive experiments affirm the efficacy and adaptability of proposed method compared to established techniques. In addition to achieving superior DG performance in terms of top-1 accuracy, proposed method demonstrates

enhanced out-of-domain calibration performance, as measured by Expected Calibration Error (ECE) and Static Calibration Error (SCE). The significance of improved calibration performance is particularly crucial in safety-critical applications, with healthcare being a notable example. In fields where prediction reliability is paramount, the proposed method holds promise for ensuring trustworthy model predictions.

REFERENCES

1. Wen R, Yuan H, Ni D, Xiao W, Wu Y. Denoising Training to Test-Time Adaptation: Enhancing Domain Generalization for Medical Image Segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision 2022*, pp 464-474.
2. Niemeijer J, Schwonberg M, Termöhlen JA, Schmidt NM, Fingscheidt T. Generalization by adaptation: diffusion-based domain extension for domain-generalized semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp 2830-2840.
3. Zhang H, Dullerud N, Seyyed-Kalantari L, Morris Q, Joshi S, Ghassemi M. An empirical framework for domain generalization in clinical settings. In *Proceedings of the Conference on Health, Inference, and Learning*, 2021, pp 279-290.
4. Bruel P, Chalamalasetti SR, Dalton CE, Hajj I, Goldman A, Graves C, Strachan JP. Generalize or die: Operating systems support for memristor-based accelerators. In *2017 IEEE International Conference on Rebooting Computing (ICRC)*, 2017, pp 1-8.
5. Ziegler H. An attempt to generalize Onsager's principle, and its significance for rheological problems. *J Applied Mathematics Physics ZAMP* 1958; 9(5-6): 748-763.
6. Doumas LA, Puebla G, Martin AE, Hummel JE. A theory of relation learning and cross-domain generalization. *Psychol Rev* 2022; 129(5): 999-1041. <https://doi.org/10.1037/rev0000346>
7. Aubreville M, Stathonikos N, Bertram CA, Klopffleisch R, Ter Hoeve N, Ciompi F, Wilm F, Marzahl C, Donovan TA, Maier A, et al. Mitosis domain generalization in histopathology images - The MIDOG challenge. *Med Image Anal* 2023; 84: 102699. doi: 10.1016/j.media.2022.102699
8. Huang X, Xiao M. Bergman-Einstein metrics, a generalization of Kerner's theorem and Stein spaces with spherical boundaries. *J Pure Applied Mathematics* 2021(770): 183-203.
9. Ghafoor R, Saleem F, Iqbal Q, Hassali MA, Hashmi FK, Haider S, Salman M. Quality of life in patients with skin diseases attending a Public Healthcare Institute of Quetta City, Pakistan. *J Pharm Pract Community Med* 2018; 4(1): 16-20
10. Hassali MA, Saleem F, Bashir S, Hashmi F, Aljadhey H. A cross-sectional assessment of health-related quality of life among type 2 diabetic patients in Pakistan. *J Pharm Bioallied Sci* 2016; 8(1): 64-68.
11. Hodzic-Hadzibegovic D, Sander BA, Lund-Andersen H. Diabetic macular oedema quantified with spectral-domain optical coherence tomography-evaluation of boundary line artefacts and the effect on retinal thickness. *Acta Ophthalmologica* 2015; 93(1): 74-82.
12. Lee JS, Kim TH, Jeon SH, Park SH, Kim SH, Lee EH, Lee JH. Automation of trimming die design inspection by the zigzag process between AI and CAD domains. *Eng Appl Artif Intell* 2024; 127: 107283.
13. Yang C, Zhu F, Liu G, Han J, Hu S. Multimodal hate speech detection via cross-domain knowledge transfer. *Proc 30th ACM Int Conf Multimedia* 2022; 1-9. <https://doi.org/10.1145/3503161.3548255>
14. Doumas LA, Puebla G, Martin AE, Hummel JE. A theory of relation learning and cross-domain generalization. *Psychological Rev* 2022; 129(5): 999-1041.
15. Awal R, Lee RK-W, Tanwar E, Garg T, Chakraborty T. Model-agnostic meta-learning for multilingual hate speech detection. *IEEE Trans Comput Soc Syst* 2023; 10(2): 1001-1012.
16. Aubreville M, Wilm F, Stathonikos N, Breininger K, Donovan TA, Jabari S, Bertram C A. A comprehensive multi-domain dataset for mitotic figure detection. *Scientific Data* 2023; 10(1): 484.
17. Meng S, Kang J, Chi K, Die X. Gearbox fault diagnosis through quantum particle swarm optimization algorithm and kernel extreme learning machine. *J Vibroengineering* 2020; 22(6): 1399-1414.
18. Jaques N, Taylor S, Sano A, Picard R. Predicting tomorrow's mood, health, and stress level using personalized multitask learning and domain adaptation. In *IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing 2017*: 17-33 PMLR.
19. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21st ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp 1721-1730.
20. Scheffer IE, Berkovic S, Capovilla G, Connolly MB, French J, Guilhoto L, Zuberi SM. ILAE classification of the epilepsies: Position paper of the ILAE Commission for Classification and Terminology. *Epilepsia* 2017; 58(4): 512-521.
21. Meuer SM, Myers CE, Klein BE, Swift MK, Huang Y, Gangaputra S, Klein R. The epidemiology of vitreoretinal interface abnormalities as detected by spectral-domain optical coherence tomography: the Beaver Dam Eye Study. *Ophthalmol* 2015; 122(4): 787-795.
22. Yang HS, Kim JT, Joe SG, Lee JY, Yoon YH. Postoperative restoration of foveal inner retinal configuration in patients with epiretinal membrane and abnormally thick inner retina. *Retina* 2015; 35(1): 111-119.

23. Arango Monnar A, Perez Rojas J, Polete Labra B. Cross-lingual hate speech detection using domain-specific word embeddings. *PLoS One* 2024; 19(7): e0299999. <https://doi.org/10.1371/journal.pone.0299999>.
24. Chiril P, Pamungkas EW, Benamara F, Moriceau V, Patti V. Emotionally informed hate speech detection: A multi-target perspective. *Cogn Comput* 2021; 13: 1312–1330. <https://doi.org/10.1007/s12559-021-09913-2>.
25. Yang H, Hua W, Xu Z, Sun J. Domain-generalized discrete diffusion model for cross-domain medical image segmentation. *IEEE Trans Med Imaging* 2025. <https://doi.org/10.1109/TMI.2025.3281234>.